

大規模言語モデルによる観光雑誌の文章校正

荒川 花蓮[†] 加勢田 健照[†] 渡部 智裕[†] 眞鍋 光汰[†] 梶原 智之[‡] 二宮 崇[‡] 後藤 功雄[‡]愛媛大学工学部工学科[†] 愛媛大学院理工学研究科[‡]

1. はじめに

観光雑誌には、地元のおすすめスポットや隠れた名所、美味しいグルメ、イベント情報など、旅行計画に役立つ実践的な情報が豊富にある。また、旅行場所を決定するきっかけにもなり、幅広い年代に利用されている。

観光雑誌を出版する際、校正することで、誤字脱字など文法的な誤りが訂正される。また、読者の印象に残るように工夫を凝らした「魅力的な表現」を届けられる。これらを人手で校正するには、多くのコストと時間が必要となる。

大規模言語モデル（LLM: Large Language Model）は膨大なデータを基に高度な言語理解と生成能力を持つため、文章校正にも活用できる。しかし、LLM に校正したい文章を入力すると、観光雑誌特有の表現が誤って修正されてしまう場合がある。

本研究では、校正内容の説明とその事例を加えたプロンプトを LLM に与えて、観光雑誌に特化した文章校正を提案する。評価実験の結果、事例だけでなく説明も付与することで、校正品質が改善できることを示す。

2. 提案手法

本研究では、LLM による観光雑誌に特化した文章校正として、事例のみだけでなく校正内容の説明も付与したプロンプトでの校正を提案する。まず 2.1 節において、校正前後の PDF から校正箇所を抽出し、データセットを構築する。次に、本データセットをもとに、説明文を作成する。そして、2.3 節においてプロンプトを構築する。

我々が取り組む LLM 校正では、校正対象となる文が 1 文ずつ入力されることを前提とする。校正する指示のみをプロンプトに与えるよりも、校正内容を説明した文と実際の校正事例を付与したプロンプトを用いる場合に、より高い校正品質が得られることが期待される。

Proofreading of Travel Magazines with Large Language Models

[†] Karen Arakawa (arakawa@ai.cs.ehime-u.ac.jp)

[†] Kensho Kaseda (h520518a@mails.cc.ehime-u.ac.jp)

[†] Tomohiro Watanabe (j520500y@mails.cc.ehime-u.ac.jp)

[†] Kota Manabe (manabe@ai.cs.ehime-u.ac.jp)

[‡] Tomoyuki Kajiwara (kajiwara@cs.ehime-u.ac.jp)

[‡] Takashi Ninomiya (ninomiya.takashi.mk@ehime-u.ac.jp)

[‡] Isao Goto (goto.isao.fn@ehime-u.ac.jp)

Ehime University

表 1: データセットの例

校正内容	校正前	校正後
表記揺れ	見た目もカワイイのでおみやげにも◎。	見た目もカワイイのでお土産にも◎。
記号	ペット用サービステラス席あり、芝生・公園で散歩可。	ペット用サービステラス席あり、芝生・公園で散歩可
誤字	西土佐のゆずを手絞りで。	西土佐のゆずを手搾りで。
表現変更	「SUKUMO モニュメント」と一緒に撮れるフォトスポットもあるので、観光の記念にぜひ！	「SUKUMO モニュメント」と一緒に撮影できるフォトスポットもあるので、観光の記念にぜひ！

2.1 データセットの構築

校正前後のデータとして、33 ペアの PDF から校正箇所を抽出したコーパスを利用する。まず、住所や休日の変更などの、情報訂正を除外する。次に、残ったデータ 112 件からランダムに抽出した 50 件のデータを、説明文の作成と、文脈内学習[1]である Few-shot に活用し、(以降、参照用データとする)残りを評価用データとする。参照用データと、文章校正の一般的な規則を参考に、抽出したコーパスの校正内容を大きく「表記揺れ」「記号」「誤字」「表現」の 4 種類にまとめる。

これにより、表 1 に示すような、校正前後のコーパスに校正内容を付与したデータセットを作成する。

2.2 説明文の作成

2.1 節のデータセットをもとに、各校正内容に対応する説明文を手で作成する。各校正内容は以下のように定義する。

- 表記揺れに関する校正 幅広い読者に利用される新聞では、中学生までに習う漢字で約 70 % から 89 % が網羅されている[2]。雑誌も新聞のように読者層が広いことから、本研究では中学生までに習う漢字に制限し、漢字と平仮名を変換する。また、数値表記の単位については国際単位系の SI 単位に揃え、SI 単位でないものは日本語の単位名称に変換する。加えて、雑誌内での表現を統一するために、「A 県 B 町」は「A 県 B」と表す。

- **記号に関する校正** 可読性向上のために情報の並列には中点を利用する．場合に依じて，句読点を追加または削除する．
- **誤字訂正** 誤字を訂正する．また，助詞の欠落および脱字を訂正する．
- **表現変更** 表現を統一するために話し言葉を書き言葉に変更する．店名および商品名といった固有名詞は「」を付けることで強調する．

これらの校正内容をまとめ，各校正に対応する説明文を作成した．

2.3 プロンプトの作成

2.1, 2.2 節から，ラベルに応じた事例を抽出し，校正内容の説明文および事例をもつプロンプトを作成する．このとき，説明文を全て付与したのち，各事例を与える．最後に，校正指示と入力文を追加する．

3. 実験設定

3.1 大規模言語モデル

本研究では，HuggingFace Transformers¹ で公開されている LLM の Swallow-8b² [3] を利用し，プロンプトエンジニアリングによる校正品質の向上を図る．

3.2 データセット

Few-shot の事例として，参照用データ 50 件から 17 件を使用する．評価には，評価用データを用いる．この時，17 件にある校正 4 種類の内訳は，表記ゆれ 5 件，句読点変更 5 件，誤字訂正 2 件，表現変更 5 件とする．

3.3 比較手法

以下の 4 手法を評価する．その際，モデルの温度パラメータを固定し，各手法の再現性を確保する．

- **ベースライン** 説明や事例を与えず，校正指示のみを与える．
- **説明のみ (A)** 校正指示に加え，説明文を与える．その際，指示文の最初に日本語の観光雑誌を取り扱うことを明記する．
- **事例のみ (B)** 各校正を含んだ 17 件の事例を校正指示とともに与える．各事例は，校正の種類ごとにまとめて与える．

¹ <https://huggingface.co/docs/transformers>

² <https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2>

表 2：実験結果

	GLEU	BLEU
ベースライン	0.330	56.97
説明のみ	0.341	58.98
事例のみ	0.361	64.55
説明 + 事例	0.384	66.08

- **説明 + 事例 (A + B)** 校正指示に加え，校正内容の説明文と各事例をまとめて与える．その際，説明文を全て付与したのち校正事例を与える．

これらの手法において，手法 A, B, および A + B は提案手法となる．各手法で得られる出力文に対して MeCab で単語分割したのち，GLEU と BLEU を用いて評価する．

4. 実験結果

表 2 に実験結果を示す．3 つの提案手法のうち，説明 + 事例が最高性能を達成した．説明と事例を与えたプロンプトを用いることで，校正品質を改善できることがわかった．また，ベースラインと比較すると，説明よりも事例を与える方が効果的だが，事例と説明を合わせることで，より高い効果が得られた．

5. おわりに

本研究では，参照用データを参考に，観光雑誌を対象とした校正の説明文とその事例を付与したプロンプトでの LLM 校正に取り組んだ．校正品質を評価した結果，事例と校正に関する説明文を付与する手法が最も有効であった．

謝辞 株式会社アイクコーポレーション，lanitech 合同会社からデータを提供いただきました．

参考文献

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et al. Language Models are Few-Shot Learners. NeurIPS, pp.1877-1901, 2020.
- [2] 野崎造成, 清水康敬. 新聞における漢字頻度特性の分析と NIE のための漢字学習表の開発, 日本教育工学会論文誌, 2000.
- [3] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities, 2024.